This Page Is Inserted by IFW Operations
and is not a part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

Freeform Search

Database:
US Pre-Grant Publication Full-Text Database
US Patents Full-Text Database
US OCR Full-Text Database
EPO Abstracts Database
JPO Abstracts Database
Derwent World Patents Index
IBM Technical Disclosure Bulletins

Term:

Display: 10 **Documents in Display Format:** - **Starting with Number** 1

Generate: ○ **Hit List** ● **Hit Count** ○ **Side by Side** ○ **Image**

Search Clear Interrupt

Search History

DATE: Tuesday, June 29, 2004 Printable Copy Create Case

Set Name **side by side**	**Query**	**Hit Count**	**Set Name** **result set**
	DB=PGPB,USPT,USOC,EPAB,JPAB,DWPI,TDBD; PLUR=YES; OP=OR		
L21	l5 and l12	46	L21
L20	l7 and l12	3	L20
L19	707/7	1568	L19
L18	707/4	3742	L18
L17	707/2	4011	L17
L16	382/229	865	L16
L15	382.clas.	42000	L15
L14	705/5	800	L14
L13	705.clas.	27593	L13
L12	707.clas.	21202	L12
L11	707/3	6618	L11
L10	707/1	6616	L10
L9	L8 and comput$	4	L9
L8	L7 and query	4	L8
L7	(on-line near analytical near mining or olam)	137	L7
L6	(intelliminer or intelli with miner)	2	L6

L5	L4 and (attribute-value$ or attribute with value$)	57	L5
L4	L3 and list	184	L4
L3	L2 and query	241	L3
L2	L1 and (datamining or data with mining)	286	L2
L1	(on-line near analytical near process$ or olap)	1052	L1

END OF SEARCH HISTORY

Refine Search

Search Results -

Terms	Documents
6044366.uref.	9

Database:

US Pre-Grant Publication Full-Text Database
US Patents Full-Text Database
US OCR Full-Text Database
EPO Abstracts Database
JPO Abstracts Database
Derwent World Patents Index
IBM Technical Disclosure Bulletins

Search: L7

Refine Search

Recall Text Clear Interrupt

Search History

DATE: Tuesday, June 29, 2004 Printable Copy Create Case

Set Name side by side	**Query**	**Hit Count**	**Set Name** result set
DB=PGPB,USPT,USOC,EPAB,JPAB,DWPI,TDBD; PLUR=YES; OP=OR			
L7	6044366.uref.	9	L7
L6	5767854.uref.	25	L6
L5	5767854.uref.	25	L5
L4	6044366.uref.	9	L4
L3	5767854.pn.	2	L3
L2	4490811.pn.	2	L2
L1	6044366.pn.	2	L1

END OF SEARCH HISTORY

Refine Search

Search Results -

Terms	Documents
5918232.pn.	2

Database: US Pre-Grant Publication Full-Text Database
US Patents Full-Text Database
US OCR Full-Text Database
EPO Abstracts Database
JPO Abstracts Database
Derwent World Patents Index
IBM Technical Disclosure Bulletins

Search:

Refine Search

Recall Text

Clear

Interrupt

Search History

DATE: Tuesday, June 29, 2004 Printable Copy Create Case

Set Name side by side	Query	Hit Count	Set Name result set
	DB=PGPB,USPT,USOC,EPAB,JPAB,DWPI,TDBD; PLUR=YES; OP=OR		
L24	5918232.pn.	2	L24
L23	5584024.pn.	2	L23
L22	6119120.pn.	2	L22
L21	(on-line near analytical near process$ or olap)	1052	L21
L20	on-line near analytical near mining	3	L20
L19	L18 and comput$	276	L19
L18	L17 and attribute near values	276	L18
L17	L16 and query$	1767	L17
L16	(intelliminer or intelli-miner or data near mining)	3873	L16
L15	(intelliminer or intelli-miner or intelligent near min$)	0	L15
L14	6539391.uref.	0	L14
	DB=USPT; PLUR=YES; OP=OR		
L13	5148379.pn.	1	L13
L12	5159687.pn.	1	L12

L11	5235701.pn.	1	L11
L10	5408638.pn.	1	L10
L9	4482971.pn.	1	L9
L8	4918643.pn.	1	L8
L7	5408638.pn.	1	L7
L6	5408638.pn.	1	L6
L5	6012058.pn.	1	L5
L4	6115708.pn.	1	L4
L3	6192360.pn.	1	L3
L2	6260036.pn.	1	L2
DB=PGPB,USPT,USOC,EPAB,JPAB,DWPI,TDBD; PLUR=YES; OP=OR			
L1	6260036.uref.	1	L1

END OF SEARCH HISTORY

Hit List

Clear | Generate Collection | Print | Fwd Refs | Bkwd Refs | Generate OACS

Search Results - Record(s) 1 through 1 of 1 returned.

□ 1. Document ID: US 6539391 B1

Using default format because multiple data bases are involved.

```
   L1: Entry 1 of 1                      File: USPT                Mar 25, 2003

US-PAT-NO: 6539391
DOCUMENT-IDENTIFIER: US 6539391 B1

TITLE: Method and system for squashing a large data set

DATE-ISSUED: March 25, 2003

INVENTOR-INFORMATION:
```

NAME	CITY	STATE	ZIP CODE	COUNTRY
DuMouchel; William H.	Chatham	NJ		
Volinsky; Christopher T.	Morris Plains	NJ		
Johnson; Theodore J.	New York	NY		
Cortes; Corinna	New York	NY		
Pregibon; Daryl	Summit	NJ		

```
US-CL-CURRENT: 707/101; 707/100, 707/102, 707/2
```

Full | Title | Citation | Front | Review | Classification | Date | Reference | [illegible] | [illegible] | Claims | KWIC | Draw De

Clear | Generate Collection | Print | Fwd Refs | Bkwd Refs | Generate OACS

Terms	Documents
6260036.uref.	1

Display Format: - Change Format

Previous Page Next Page Go to Doc#

First Hit Fwd Refs

□ Generate Collection Print

L5: Entry 53 of 57 File: USPT Dec 14, 1999

DOCUMENT-IDENTIFIER: US 6003029 A
TITLE: Automatic subspace clustering of high dimensional data for data mining applications

Parent Case Text (2):
The present application is related to an application entitled "Discovery-Driven Exploration Of OLAP Data Cubes," by Sunita Sarawagi and Rakesh Agrawal, Ser. No. 08/916,346 filed on Aug. 22, 1997, having common ownership, filed concurrently with the present application, and incorporated by reference herein.

Brief Summary Text (3):
The present invention relates to the field of computing. More particularly, the present invention relates to an approach for organizing data within a dataset for data mining.

Brief Summary Text (5):
Clustering is a descriptive task associated with data mining that identifies homogeneous groups of objects in a dataset. Clustering techniques have been studied extensively in statistics, pattern recognition, and machine learning. Examples of clustering applications include customer segmentation for database marketing, identification of sub-categories of spectra from the database of infra-red sky measurements, and identification of areas of similar land use in an earth observation database.

Brief Summary Text (6):
Clustering techniques can be broadly classified into partitional techniques and hierarchial techniques. Partitional clustering partitions a set of objects into K clusters such that the objects in each cluster are more similar to each other than to objects in different clusters. For partitional clustering, the value of K can be specified by a user, and a clustering criterion must be adopted, such as a mean square error criterion, like that disclosed by P. H. Sneath et al., Numerical Taxonomy, Freeman, 1973. Popular K-means methods, such as the FastClust in SAS Manual, 1995, from the SAS Institute, iteratively determine K representatives that minimize the clustering criterion and assign each object to a cluster having its representative closest to the cluster. Enhancements to partitional clustering approach for working on large databases have been developed, such as CLARANS, as disclosed by R. T. Ng et al., Efficient and effective clustering methods for spatial data mining, Proc. of the VLDB Conference, Santiago, Chile, September 1994; Focussed CLARANS, as disclosed by M. Ester et al., A database interface for clustering in large spatial databases, Proc. of the 1st Int'l Conference on Knowledge Discovery in Databases and Data Mining, Montreal, Canada, August 1995; and BIRCH, as disclosed by T. Zhang et al., BIRCH: An efficient data clustering method for very large databases, Proc. of the ACM SIGMOD Conference on Management Data, Montreal, Canada, June 1996.

Brief Summary Text (7):
Hierarchial clustering is a nested sequence of partitions. An agglomerative, hierarchial clustering starts by placing each object in its own atomic cluster and then merges the atomic clusters into larger and larger clusters until all objects are in a single cluster. Divisive, hierarchial clustering reverses the process by

starting with all objects in cluster and subdividing into smaller pieces. For theoretical and empirical comparisons of hierarchical clustering techniques, see for example, A. K. Jain et al., Algorithms for Clustering Data, Prentice Hall, 1988, P. Mangiameli et al., Comparison of some neutral network and hierarchical clustering methods, European Journal of Operational Research, 93(2):402-417, September 1996, P. Michaud, Four clustering techniques, FGCS Journal, Special Issue on Data Mining, 1997, and M. Zait et al., A Comparative study of clustering methods, FGCS Journal, Special Issue on Data Mining, 1997.

Brief Summary Text (8):
Emerging data mining applications place special requirements on clustering techniques, such as the ability to handle high dimensionality, assimilation of cluster descriptions by users, description minimation, and scalability and usability. Regarding high dimensionality of data clustering, an object typically has dozens of attributes in which the domains of the attributes are large. Clusters formed in a high-dimensional data space are not likely to be meaningful clusters because the expected average density of points anywhere in the high-dimensional data space is low. The requirement for high dimensionality in a data mining application is conventionally addressed by requiring a user to specify the subspace for cluster analysis. For example, the IBM data mining product, Intelligent Miner described in the IBM Intelligent Miner User's Guide, version 1 release 1, SH12-6213-00 edition, July 1996, and incorporated by reference herein, allows specification of "active" attributes for defining a subspace in which clusters are found. This approach is effective when a user can correctly identify appropriate attributes for clustering.

Brief Summary Text (9):
A variety of approaches for reducing dimensionality of a data space have been developed. Classical statistical techniques include principal component analysis and factor analysis, both of which reduce dimensionality by forming linear combinations of features. For example, see R. O. Duda et al., Pattern Classification and Scene Analysis, John Wiley and Sons, 1973, and K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, 1990. For the principal component analysis technique, also known as Karhunen-Loeve expansion, a lower-dimensional representation is found that accounts for the variance of the attributes, whereas the factor analysis technique finds a representation that accounts for the correlations among the attributes. For an evaluation of different feature selection methods, primarily for image classification, see A. Jain et al., Algorithms for feature selection: An evaluation, Technical report, Department of Computer Science, Michigan State University, East Lansing, Mich., 1996. Unfortunately, dimensionality reductions obtained using these conventional approaches conflict with the requirements placed on the assimilation aspects of data mining.

Brief Summary Text (10):
Data mining applications often require cluster descriptions that can be assimilated and used by users because insight and explanations are the primary purpose for data mining. For example, see U. M. Fayyad et al., Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, 1996. Clusters having decision surfaces that are axis parallel and, hence, can be described as Disjunctive Normal Form (DNF) expressions, become particularly attractive for user assimilation. Nevertheless, even while a description is a DNF expression, there are clusters that are poorly approximated poorly, such as a cigar-shaped cluster when the cluster description is restricted to be a rectangular box. On the other hand, the same criticism can also be raised against decision-tree and decision-rule classifiers, such as disclosed by S. M. Weiss et al., Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems, Morgan Kaufman, 1991. However, in practice, the classifiers exhibit competitive accuracies when compared to techniques, such as neural nets, that generate considerably more complex decision surfaces, as disclosed by D. Michie, Machine Learning, Neural and

Statistical Classification, Ellis Horwood, 1994.

Brief Summary Text (20):
The foregoing clustering model can be considered nonparametric in that mathematical forms are neither assumed for data distribution, nor for clustering criteria Instead, data points are separated according to the valleys of a density function, such as disclosed by K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, 1990. One example of a density-based approach to clustering is DBSCAN, as disclosed by M. Ester et al., A density-based algorithm for discovering clusters in large spatial databases with noise, Proc. of the 2nd Int'l Conference on Knowledge Discovery in Databases and Data Mining, Portland, Oregon, August 1995. This approach defines a cluster as a maximal set of density-connected points. However, the application domain for DBSCAN is spatial databases and with interest in finding arbitrarily-shaped clusters.

Brief Summary Text (21):
Several other techniques for nonparametric clustering are based on estimating density gradient for identifying valleys in the density function. These techniques are computationally expensive and generally result in complex cluster boundaries, but which may provide the most correct approach for certain data mining applications.

Brief Summary Text (26):
The present invention automatically identifies subspaces in a multi-dimensional data space in which clusters are found and provides description assimilation for a user, description minimization and scales as the size of the data space increases. The present invention finds clusters embedded in subspaces of high-dimensional data without requiring a user to guess subspaces that might have interesting clusters. The cluster descriptions generated by the present invention are DNF expressions having minimal descriptions for ease of user comprehension. The present invention is insensitive to the order of records input The present invention is a basic data mining operation along with other operations, such as associations and sequential-patterns discovery, time-series clustering, and classification.

Detailed Description Text (5):
For the first phase of the present invention, the simplest way for identifying dense units would be to create a histogram in all subspaces and count the points contained in each unit during one pass over the data. However, this approach is infeasible for high dimensional data. Consequently, the present invention uses a bottom-up cluster identification approach that is similar to the conventional Apriori algorithm for finding Association rules, as disclosed by R Agrawal et al., Fast Discovery of Association Rules, Advances in Knowledge Discovery and Data Mining, U. M. Fayyad et al., editors, AAAI/MIT Press, Chapter 12, pages 307-328, 1996, and incorporated by reference herein. A similar bottom-up cluster identification approach for determining modes in high-dimensional histograms is disclosed by R. S. Chhikara et al., Register A numerical classification method for partitioning of a large multidimensional mixed data set, Technometrics, 21:531-537, 1979.

Detailed Description Text (14):
For the present invention, the sorted list of subspaces are divided into two clusters, a pruned set and a live set. The model used by the present invention is for each set, the mean is kept, and for each subspace, the deviation from the mean is kept. The code length is the sum of the bit lengths of the numbers needed to be stored For example, assume the subspaces S.sub.1, S.sub.2, . . . , S.sub.n, and the sorted list of the covered fractions being respectively x.sub.S1, x.sub.S2, . . . , x.sub.Sn. Then, if dimensions x.sub.S1+1, . . . , x.sub.Sn are decided to be pruned, then the length of the encoding is: ##EQU2##

Detailed Description Text (22):

The asymptotic running times of the present invention are described in terms of dense units accesses. Dense units are stored in a data structure, e.g., a hash tree, that allows efficient queries. For each maximal region R, the greedy-growth approach performs O(Size(R)) dense unit accesses, where Size(R) is the number of dense units contained in R. This is shown by letting S be the subspace that R lies in, K the number of dimensions of S, and N the number of dense units in S. The greedy-growth approach accesses each unit that R covers for ascertaining whether R is indeed part of a cluster. Additionally, the greedy-growth approach also accesses each neighbor unit of R for ascertaining whether R is maximal and, consequently, no region that is a proper superset of R is part of a cluster. However, the number of neighbor units is bound by 2kSize(R).

Detailed Description Text (25):
The sequential-scan approach of the present invention that determines maximal regions that cover a cluster computes all maximal regions that cover all the dense units lying in a given subspace. The main idea of the sequential-scan approach is to apply an efficient dynamic programming technique. Let S be a subspace having dimensions a.sub.1, a.sub.2, . . . , a.sub.k. A total ordering of the dense units is imposed, and the units are visited according to that order. For each unit d, all maximal regions are found that have d as their upper right comer. Since every maximal region has a unit in its corner, all maximal regions are enumerated in this way. The list of maximal regions for each unit is computed using local information. In particular, for each unit, K different sets of regions are computed, M.sub.i (d),1.ltoreq.i.ltoreq.k. M.sub.i (d) is the set of maximal regions when S is projected to the first i dimensions (a1, . . . , ai), and to units that appear lower than d in these dimensions. Every region in M.sub.i (d) either covers the previous unit d.sub.prev on the i-th dimension, or it also belongs to M.sub.i-1 (d). To compute M.sub.i (d), M.sub.i-1 (d) and M.sub.i (d.sub.prev) can be used. Once all M.sub.i (d) are computed, the regions in M.sub.k (d), .A-inverted.d.epsilon.S are examined for finding which regions are maximal.

Detailed Description Text (36):
The Removal Heuristic, on the other hand, is easy to implement and efficient in execution. It needs a simple scan of the sorted list of regions requiring exactly Size(R) dense unit accesses for each region. The total number of accesses is then .SIGMA.Size(R.sub.i)=O(n.sup.2). A disadvantage of the Removal Heuristic comes from the worse-case results for the set-cover problem.

Detailed Description Text (39):
A modified version of the synthetic data generation program used by M. Zait et al., A Comparative study of clustering methods, FGCS Journal Special Issue on Data Mining, 1997, and incorporated by reference herein, was used for a comparative study of conventional clustering algorithms. The data generation program produced datasets having clusters of high density in specific subspaces and provided parameters for controlling the structure and the size of datasets, such as the number of records, the number of attributes, and the range of values for each attribute. A bounded data space (n-dimensional cube) that data points live in was assumed. Each data space was partitioned into a multi-dimensional grid generated by an equi-width partitioning of each dimension into 10 partitions. Each box of the grid formed a unit.

Detailed Description Text (53):
FIG. 9 shows a program storage device 90 having a storage area 91. Information stored in the storage area in a well-known manner that is readable by a machine, and that tangibly embodies a program of instructions executable by the machine for performing the method of the present invention described herein for automatically finding subspaces of the highest dimensionality in a data space for data mining applications. Program storage device 90 can be a magnetically recordable medium device, such as a magnetic diskette or hard drive, or an optically recordable medium device, such as an optical disk.

Detailed Description Text (54):
Although the present invention is primarily focused on data mining applications, the techniques of present invention are also applicable to OLAP databases. To index OLAP data, for instance, a data space is first partitioned into dense and sparse regions, as disclosed by U.S. Pat. No. 5,359,724 to R. J. Earle. Data in dense regions is stored in an array, whereas a tree structure is used for storing sparse regions. Currently, users are required to specify dense and sparse dimensions. Similarly, the precomputation techniques for range queries over OLAP data cubes require identification of dense regions in sparse data cubes. The present invention can be used for identification of dense regions in sparse data cubes.

Other Reference Publication (1):
R. Agrawal et al., Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications, Paper AR.sub.-- 297, pp. 1-18, 1997.

Other Reference Publication (4):
A. Arning et al., A Linear Method for Deviation Detection in Large Databases, Proceedings of the 2nd International Conference on Knowledge Discovery in Databases and Data Mining, pp. 164-169, Portland, Oregon, Aug., 1996.

Other Reference Publication (5):
C.J. Matheus et al., Selecting and Reporting What is Interesting, Advances in Knowledge Discovery and Data Mining, pp. 495-515, AAAI Press, 1996.

Other Reference Publication (10):
R. Agrawal et al. Database Mining: A Performance Perspective, IEEE Transactions on Knowledge and Data Engineering, vol. 5, No. 6, Dec. 1993, pp. 914-925.

First Hit Fwd Refs

End of Result Set

□ Generate Collection Print

L6: Entry 2 of 2 File: USPT Jul 10, 2001

DOCUMENT-IDENTIFIER: US 6260036 B1

**** See image for Certificate of Correction ****

TITLE: Scalable parallel algorithm for self-organizing maps with applications to sparse data mining problems

Detailed Description Text (94):

In this paper, we have developed a data-partitioned parallel method for the well-known self-organizing feature map developed by Kohonen. Our approach is based on an enhanced version of the batch SOM algorithm which is particularly efficient for sparse data sets encountered in retail data mining studies. We have demonstrated the computational efficiency and parallel scalability of this method for sparse and non-sparse data, using 3 data sets, two of which include actual retail spending data. Model problem analysis, plus visualizations of the segmentations produced for publicly available census data have shown that the batch SOM methodology provides reasonable clustering results and useful insights for data mining studies. Algorithms similar to those discussed in this paper are planned for inclusion in a future release of the IBM Intelligent Miner [IBM Intelligent Miner, http://www.software.ibm.com/data/intelli-mine] data mining product.